Claude 3 Opus 和 Kimi 在生物医学期刊同行评议中的能力评价[1]

倪明[1, 2)]

1）复旦大学附属肿瘤医院杂志社办公室，复旦大学上海医学院肿瘤学系，上海市徐汇区东安路 270 号 200032

2）《中国癌症杂志》杂志社有限公司，上海市徐汇区东安路270号 200032

摘 要：**[目的]** 本研究旨在评估 Claude 3 Opus 和 Kimi 在生物医学期刊同行评议中的能力。**[方法]** 对《中国癌症杂志》进入定稿会终审的 29 篇论文，采用 Claude 3 Opus 和 Kimi 进行同行评议，并根据生物医学研究报告指南中的披露清单进行审查，所有论文作者均授权同意用 AIGC 进行同行评议。采用李克特 5 分量表对 2 者同行评议结果进行评分。计数资料采用单因素方差检验或 *Fisher* 精确概率检验，多组配对计量资料采用 Friedman *M* 检验，四格表法计算 2 者评议结果的灵敏度、特异度、阳性预测值、阴性预测值和准确性，并绘制 ROC 曲线，评估其预测能力。**[结果]**29 篇文章，定稿会评议后发表 6 篇、修改后发表 15 篇、退稿 8 篇，Claude 3 Opus 同行评议结论发表 19 篇、修改后发表 10 篇，Kimi 同行评议结论发表 9 篇、修后发表 16 篇、退稿 4 篇，Friedman *M* 检验结果显示，Kimi 与定稿会专家评议结论差异无统计学意义（*M*=0.241，调整后 *P*=1.000）。李克特量表结果显示：专家对 Kimi 评议结果认可度高于 Claude 3 Opus 的评议结果（3.85±0.47 *vs* 3.48±0.73，*F*=10.017，*P*=0.002）。在生物医学研究报告指南披露清单的审查方面，Claude 3 Opus 评价的准确性为 77.5%，灵敏度为 76.9%，特异度为 64.0%；Kimi 评价的准确性为 75.2%，灵敏度为 77.5%，特异度为 70.1%。ROC 曲线分析结果显示，Claude 3 Opus 和 Kimi 曲线下面积分别为 0.818 和 0.841，有较好的预测能力。经检验，Kimi 的审查结果与责编审查结果差异无统计学意义（*M*=-0.152，调整后 *P*=0.061）。**[结论]** Claude 3 Opus 和 Kimi 在生物医学研究报告指南披露清单的审查方面表现出较好的能力，与责任编辑审查有较高的一致性。然而两款 AIGC 模型尚未能达到专家同行评议的能力，存在①生成内容不准确性；②生成内容缺乏个性化；③外部内容的推演不足；④评审内容的模糊性；⑤生成内容过于粗况；⑥审稿结论偏于正向评价等问题，但是在展现出一定应用潜力。为了进一步提高其效能，应开发高质量的 AIGC 专用工具辅助用于生物医学期刊的同行评议，建议医学专家、医学编辑和 AIGC 开发者共同努力，制定相关标准，确保数据安全和质量，提高透明度，减少评审偏倚，遵循出版伦理，并建立有效的监督和反馈机制，以确保 AIGC 在生物医学期刊同行评议中的准确性。

关键词：生物医学期刊；生成式人工智能；Claude 3 Opus；Kimi；同行评议

2022 年 11 月，美国 OpenAI 公司发布了全新聊天机器人模型 ChatGPT，引起了轰动，发布后首月的用户达到了 5 700 万，2024 年 2 月，又发布了 ChatGPT-4，标志着 AI 在理解和生成自然语言方面取得了重大进展[1-2]。因 AIGC 在人机交互的便利性和智能性，在工作效率的提升方面展示出巨大的能力，期刊界也在尝试 AIGC 应用于优化编辑出版工作，包括辅助写作、文献综述、内容翻译、编辑校对、专刊策划、自动摘要、学术内容科普化、封面制作、学术不端检测、传播营运，甚至配套文章视频的生成等[3-12]。2024 年，Liang 等[13]的一项研究发布至预印本 arXiv 平台上，并被 *Nature* 杂志引以新闻形式报道：同行评议人员正在采用 ChatGPT 和其他 AI 工具来评估他人的工作[14]。引发了 AIGC 能否应用于学术论文的同行评议这一热点且争议的问题。随着生物医学论文投稿量、发表量的逐年上升，加上审稿专家临床科研业务繁重，审稿超时、敷衍甚至拒绝审稿较为常见，延长论文出版时间，影响了期刊出版流程，所以不管审稿专家还是期刊界迫切地想得到肯定的回答。这

[1] 作者简介：倪明（ORCID: 0000-0001-6024-9658），硕士，副编审， E-mail: niming@shca.org.cn。

个问题前期的研究以问卷调研为主，支持的观点主要集中在①提升审稿效率；②提供无偏见的评议；③提取数据和结构化优势等[15-19]。但也提出了担忧，如学术伦理问题、版权争议问题、技术细节缺乏和隐私问题等[15]。也研究尝试将AIGC应用于生物医学论文同行评议中，如Kadi等[20]采用ChatGPT4.0对15份已经发表的病例报告进行了同行评议，表示其评审能力不如文本生成能力，且无法识别出那些已经过重大内容更改的文章中的主要不一致之处。Yorukoglu[21]采用ChatGPT和Google Bard两个AI模型对1篇文章进行了审稿，并与专家审稿相比，在数据提取、结构化分析方面表现出了创新性和实用性，有潜力提高审稿效率。但两个研究所选择的文章类型单一，均为病例报告，而且所研究的文章数量偏少。

随着AIGC不断地迭代更新，算法更优、能力更强、性能更高、安全更好、理解力更自然的AIGC模型不断涌现，更强大的人工智能模型是否在论文同行评议中取得了一定进展？Claude 3 Opus模型是由Anthropic公司开发，于2024年3月4日发布，在处理复杂任务和理解上下文方面表现得更为出色，能够更好地理解和产生更自然的语言，而且避免了生成有害的或者不恰当的内容，在安全性方面有一定优势[22]，在Elo上的排名超过了GPT4+，排名第一[23]。Kimi模型于2023年10月发布，主要应用于信息检索、数据分析、语言翻译、编程辅助以及日常对话等，具有快速处理和分析大量信息，同时保持对话的流畅性和准确性的优势，用户访问量已排名第三，且能够处理200万字的超长文本[24]。生物医学论文的写作和发表有其特有的格式，如有比较成熟和公认的生物医学研究报告指南（以下称简称《报告指南》）[23]，本研究采用上述2中AIGC模型对不同类型的生物医学论文进行同行评议，并对按照《报告指南》的条目清单来审查论文的披露情况，以评估它们在生物医学期刊同行评议中的能力和潜在应用价值。

1 研究设计

1.1 资料来源

选取2024年3月27日—5月15日间，通过《中国癌症杂志》两轮同行评议并第一轮修回的文章。文章入选标准：①同行评议意见为“修改后再审”“修改后发表”或“发表”；②文章已根据上述两轮同行评议意见修改，并于2024年5月22日前修回；③文章作者同意本刊采用AIGC模型对其修改后的文章再次进行同行评议。文章排除标准：①未通过初审、学术初审或者同行评议的；②未按时、按要求修回的；③作者不同意授权本刊采用AIGC模型进行同行评议的。

1.2 研究方法

1.2.1 文章分类和报告规范选择

根据纳入文章类型不同，将文章分为指南共识类文章、临床试验类文章、观察性研究类文章、临床预测研究类文章、诊断研究类文章、动物实验类文章、综述类文章、病例报告类文章八大类型。根据不同的文章类型选择不同的《报告指南》来评估文章写作的规范性、内容披露的完整性和研究透明度（见表1）。

表1 文章分类及对应的生物医学报告规范

文章分类	报告规范	报告指南网址	披露清单
指南共识	AGREE Ⅱ	https://www.agreetrust.org/	23条
临床试验	CONSORT	http://www.consort-statement.org/	25条
观察性研究	STROBE	https://www.strobe-statement.org/	22条
临床预测研究	TRIPOD	https:/www.tripod-statement.org/	27条
诊断研究	STARD	https://www.equator-network.org/reporting-guidelines/stard/	30条
动物实验	ARRIVE	https://arriveguidelines.org/	21条
综述	PRISMA	http://www.prisma-statement.org/	27条
病例报告	CARE	https://www.care-statement.org/	13条

1.2.2 AIGC模型的注册及训练

①Claude 3 Opus 模型：将 Sider 智能工具（https://sider.ai/）嵌入到 Edge 浏览器插件中，注册并成为付费用户。Sider 集成了多种高级 AI 模型，如 GPT-4o、GPT-3.5、Claude 3 和 Gemini 1.5 等，能实现聊天交互、快速阅读与总结、改善写作质量等，本研究选择最新的 Glaude 3 Opus 模型。

②Kimi 模型：登录 Kimi 官网(https://kimi.moonshot.cn)，并根据提示完成注册。

根据《中国癌症杂志》对稿件同行评议要求，对两种 AIGC 进行提问训练，最后形成 2 个提示词（见表 2），提示词 1 主要针对论文的同行评议，提示词 2 主要审查论文中《报告指南》披露清单披露情况。在采用提示词 2 进行提问之前，分别向 2 种 AIGC 模型上传《报告指南》的披露清单，并要求 2 种 AIGC 模型学习并完整地提取披露清单（提示词句为：请学习并完整地翻译上传附件中的 XXX 披露清单），再让 AIGC 模型回答提示词 2 问题。同一类型的文章必须在同一个对话中完成，不同类型文章，开启新的对话。

表 2 向 2 种 AIGC 模型提问的提示词

文章类型	**提示词 1（提问时间 2024 年 5 月 20 日）**	**提示词 2（提问时间 2024 年 6 月 11-28 日）**
指南共识	请从文章内容的新颖性，撰写专家的权威性，参考文献的权威性，对读者的启发方面进行同行评议，明确提出文章的优点和不足，并对该文是否达到《中国癌症杂志》的发表水平给出意见。	请按照所学的 AGREE Ⅱ中的 23 条披露清单，对此文章进行评价，并按照披露清单，逐条判断文章是否完整地进行披露，并写出披露与否的原因
临床试验/观察性研究/临床预测研究/诊断类研究/动物实验	请从文章内容的创新性，数据处理的可靠性，研究设计的科学性，研究结果是否有临床应用价值，以及是否有类似的文章已经发表等方面进行同行评议，明确提出文章的优点和不足，并对该文是否达到《中国癌症杂志》的发表水平给出意见。	请按照所学的 XXX 标准中的 XX 条披露清单，对此文章进行评价，并按照披露清单，逐条判断文章是否完整地进行披露，并写出披露与否的原因
综述	请从文章内容的新颖性，撰写专家的权威性，参考文献的权威性，对读者的启发方面进行同行评议，明确提出文章的优点和不足，并对该文是否达到《中国癌症杂志》的发表水平给出意见。	请按照所学的 PRISMA 中的 27 条披露清单，对此文章进行评价，并按照披露清单，逐条判断文章是否完整地进行披露，并写出披露与否的原因。如果该文不是 meta 类文章，不适用 PRISMA 进行评价，如何让作者按照 PRISMA 要求进行修改？
病例报告	请从病例信息描述的完整性，病例的罕见性，对读者的启发方面进行同行评议，明确提出文章的优点和不足，并对该文是否达到《中国癌症杂志》的发表水平给出意见。	请按照所学的 CARE 声明中的 13 条披露清单，对此文章进行评价，并按照披露清单，逐条判断文章是否完整地进行披露，并写出披露与否的原因

1.2.3 AIGC 模型生成内容的评价

针对 2 种 AIGC 模型对提示词 1 生成的内容，由责任编辑进行整理，并设计问卷，内容包含本研究背景内容、对 AIGC 所提的问题和 AIGC 的评议内容，并明确提示专家按照《中国癌症杂志》同行评议要求，对文章再次进行独立审稿，并对 AIGC 生成的同行评议内容采取李克特 5 分量表法进行评价（见表 3）。定稿会专家评价完所有 AIGC 生成的同行评议报告以后，对 AIGC 在生物医学期刊同行评议中是否有具有应用前景进行整体评价。针对问题 2 生成的内容，由 2 名副编审通读文章后，按照对应《报告指南》的披露清单进行逐条核对，如 2 者意见不一致，重新阅读文章，商讨后直到形成一致意见为止。将副编审评价的结果作为金标准与 AIGC 的评价结果进行比较和统计分析。

表 3 李克特 5 分量表评价标准

专家意见	赋值	解释
非常同意	5	完全同意 Glaude 3 Opus 或 Kimi 生成的审稿意见
同意	4	一定程度同意 Glaude 3 Opus 或 Kimi 生成的审稿意见
一般	3	对 Glaude 3 Opus 或 Kimi 生成的审稿意见持中立态度

不同意	2	一定程度不同意 Glaude 3 Opus 或 Kimi 生成的审稿意见
非常不同意	1	非常不同意 Glaude 3 Opus 或 Kimi 生成的审稿意见

1.2.4 统计分析

采用 SPSS 22 对数据进行统计分析，计数资料采用皮尔森卡方检验或者 *Fisher* 精确概率法检验，计量资料采用单因素方差分析，用多个配对样本 Friedman *M* 检验，采用四格表法计算灵敏度、特异度、阳性预测值和阴性预测值，绘制 ROC 曲线，来评估 AIGC 模型的预测性能。取 α=0.05 为检验水准，双侧检验。

1.3 研究思路和执行方案

本研究文章的纳入流程、研究思路和执行步骤详见图 1。

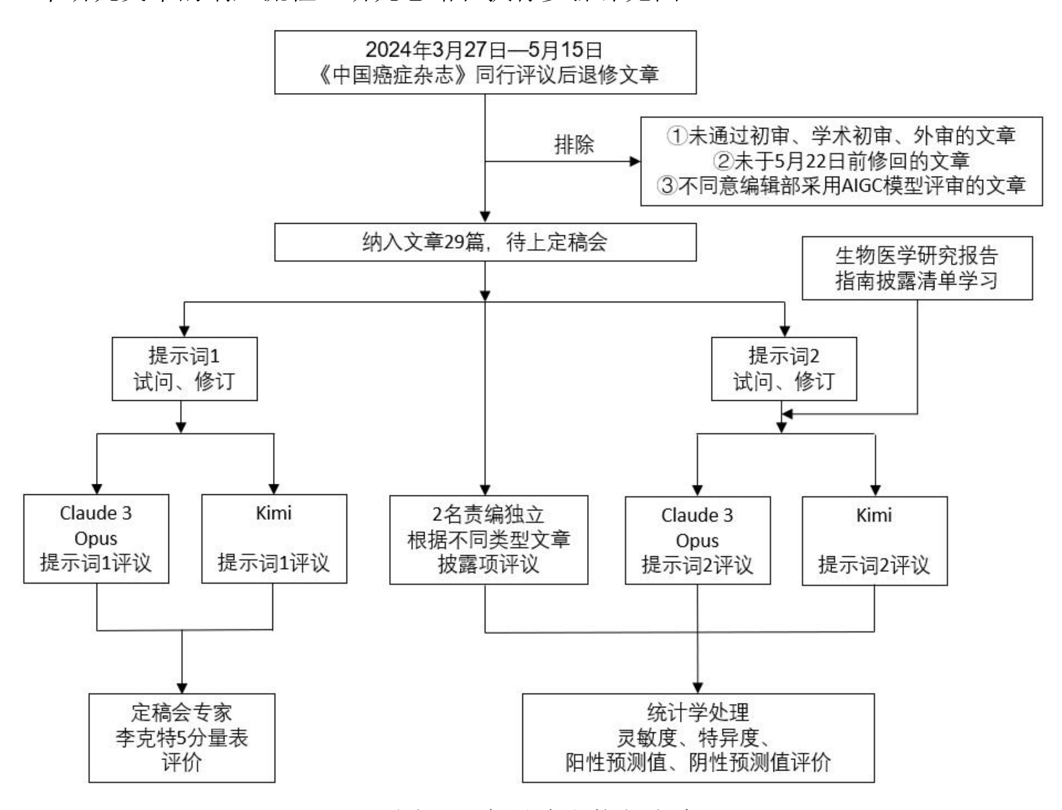


图 1 研究思路和执行方案

2. 研究结果

2.1 被研究论文的纳入情况和审稿结论

本研究共纳入符合条件的论文 29 篇，其中综述类文章 9 篇，观察性研究 8 篇，指南共识类和模型预测类文章均为 4 篇，定稿会建议直接发表 6 篇，修后发表 15 篇，退稿 8 篇，2 种 AIGC 模型均对文章是否录用都给出了审稿结论，具体见表 4。

表 4 纳入文章情况和同行评议审稿结论性建议

序号	文章类型	定稿会审稿结论	Claude 3 Opus 审稿结论	Kimi 审稿结论
1	随机对照研究	发表	发表	可发表，建议作者进一步提供长期随访数据，扩大样本量进行多中心研究，并考虑患者的生活质量等临床指标。
2	指南共识	发表	发表	适合发表
3	指南共识	修后发表	值得发表	适合发表
4	指南共识	修后发表	值得发表	适合发表
5	指南共识	修后发表	值得发表	适合发表

6	综述	发表	值得发表	适合发表
7	综述	修后发表	值得发表	适合发表
8	综述	修后发表	值得发表	适合发表
9	综述	退稿	值得发表	适合发表
10	综述	退稿	值得发表	未达到发表水平
11	综述	退稿	值得发表	未达到发表水平
12	综述	退稿	值得发表	未达到发表水平
13	综述	发表	值得发表	已具备发表条件，但需在细节上做适当改进
14	综述	发表	值得发表	已具备发表条件
15	观察性研究	修后发表	值得发表	可发表，建议作者在数据完整性、临床相关性和原创性方面进行加强
16	观察性研究	修后发表	值得发表	可发表，建议作者在数据完整性、临床相关性和原创性方面进行加强
17	观察性研究	退稿	值得发表	可发表，建议作者在数据完整性、临床相关性和原创性方面进行加强
18	观察性研究	发表	值得发表	已具备发表条件，但需在细节上做适当改进
19	观察性研究	修后发表	值得发表	已具备发表条件，但需在细节上做适当改进
20	观察性研究	退稿	值得发表。建议作者可以进一步深入探讨 IREB2-TFRC-FTH1 轴及 FBXL5 调控铁稳态的具体分子机制	未达到发表水平
21	观察性研究	退稿	值得发表。建议作者可以进一步深入探讨 AQP9 调控炎症及氧化应激的具体分子机制	建议作者进一步验证研究结果的临床相关性，并考虑增加样本量和深入探讨分子机制
22	观察性研究	修后发表	值得发表。建议作者可以进一步扩大样本量,探讨不同 PSMA PET/CT 参数与 ctDNA 特征在 HSPC 预后中的作用	可发表，建议作者进一步验证研究结果的普适性，并考虑扩大样本量和进行多中心研究
23	诊断研究	修后发表	值得发表。建议作者可以进一步扩大样本量,深入探讨影像组学特征与胸腺瘤生物学行为的关系	可发表，建议作者进一步验证模型的普适性和稳定性，考虑使用外部数据集进行验证
24	风险模型预测	修后发表	值得发表。建议作者可以进一步扩大样本量,深入探讨免疫细胞比例与肝癌预后的关系	可发表，建议作者进一步验证模型的普适性和稳定性，并考虑增加对生物学机制的探讨
25	风险模型预测	修后发表	值得发表。建议作者可以进一步扩大样本量,深入探讨影像学指标与 apCR 之间的关系	可发表，建议作者进一步验证模型的普适性，并考虑增加对外部数据集的分析
26	风险模型预测	修后发表	值得发表。建议作者可以进一步扩大样本量,深入探讨不良病理学特征数量与结直肠癌预后的关系	可发表，建议作者进一步验证模型的普适性，并考虑增加对外部数据集的分析
27	风险模型预测	修后发表	值得发表。建议作者可以进一步扩大样本量,深入探讨影像组学特征与 RFA 治疗预后的关系	可发表，建议作者进一步验证模型的普适性，并考虑增加对外部数据集的分析
28	个案报道	退稿	值得发表。建议作者可以进一步扩大病例数,并深入探讨转化机制	考虑到样本量较小且缺乏对照，未达到发表标准
29	动物实验	修后发表	值得发表，进一步探讨这些乳腺癌原代细胞系在药物筛选和基础研究中的具体应用	可发表，建议作者进一步验证细胞系的长期稳定性，并与其他细胞系进行比较分析

2.2 2 种 AIGC 模型和定稿会审稿结论之间的比较结果

根据审稿建议，将审稿结论归纳为：发表、修后发表和退稿（见表 5），其中 Claude 3

Opus 模型无一篇论文退稿。经 Friedman *M* 检验发现，Claude 3 Opus 的审稿结论与定稿会的审稿结论之间，差异有统计学意义（*M*=0.845，调整后 *P*=0.004），说明两组间审稿结论有差异；Kimi 的审稿结论与定稿会审稿结论之间，差异无统计学意义（*M*=0.241，调整后 *P*=1.000），说明两组间审稿结论无差异。

进一步将“指南共识”“综述”类文章归纳为“综述类论文”，将“随机对照研究”“观察性研究”“诊断研究”“风险模型预测”“动物实验”归纳为“研究类论文”，进行分层分析，结果显示 Claude 3 Opus 对“综述类论文”的同行评议结论与定稿会审稿结论差异有统计学意义（*M*=1.038，*P*=0.024），而 Kimi 与定稿会结论差异无统计学意义（*M*=0.577，*P*=0.424）。“论著类论文”的 Claude 3 Opus 和 Kimi 的同行评议结论与定稿会结论差异无统计学意义（*M*=0.633，*P*=0.249；*M*=-0.033，*P*=1.00）。

表 5 三者审稿结论情况 （*n*，%）

评审者	审稿结论		
	发表	修后发表	退稿
定稿会	6（20.7%）	15（51.7%）	8（27.6%）
Claude 3	19（65.5%）	10（34.5%）	0（0）
Kimi	9（31.0%）	16（55.2%）	4（13.8%）

2.3 定稿会专家对 2 种 AIGC 同行评议结果的认可度结果

李克特评分表结果显示，定稿会专家对 Kimi 的同行评议的认可度（3.85±0.47）高于 Claude 3 Opus 的认可度（3.48±0.73），经单因素方差分析检验，差异有统计学意义（F=10.017，P=0.002），但是 2 者的李克特评分均值未到 4 分，专家对 AIGC 的同行评议结论的认可度处于一般和同意之间（见表 6）。根据研究类论文和综述类论文分层结果见雷达图（图 2,3），Kimi 对研究类论文的“创新性”“文章主要不足”以及综述类论文的“文章主要有点”“文章主要不足”方面的评价均分超过 4 分，达到李克特评分中的同意级别，其余评分均为达到 4 分。其中 Claude 3 Opus 在大部分的“文章主要优点”和“文章主要不足”方面评价缺失，故得分接近为 0。11 位专家均表示 2 种 AIGC 在生物医学期刊同行评议中具有应用前景。

表 6 两种 AIGC 同行评议结果的认可度

评审者	专家评分（*n*）					x^2	*P* 值	均值±标准差	*F*	*P* 值
	非常不同意 1 分	不同意 2 分	一般 3 分	同意 4 分	非常同意 5 分					
Claude 3 Opus	75	157	717	426	304	76.141	<0.001	3.48±0.73	10.017	0.002
Kimi	60	114	909	695	607			3.85±0.47		

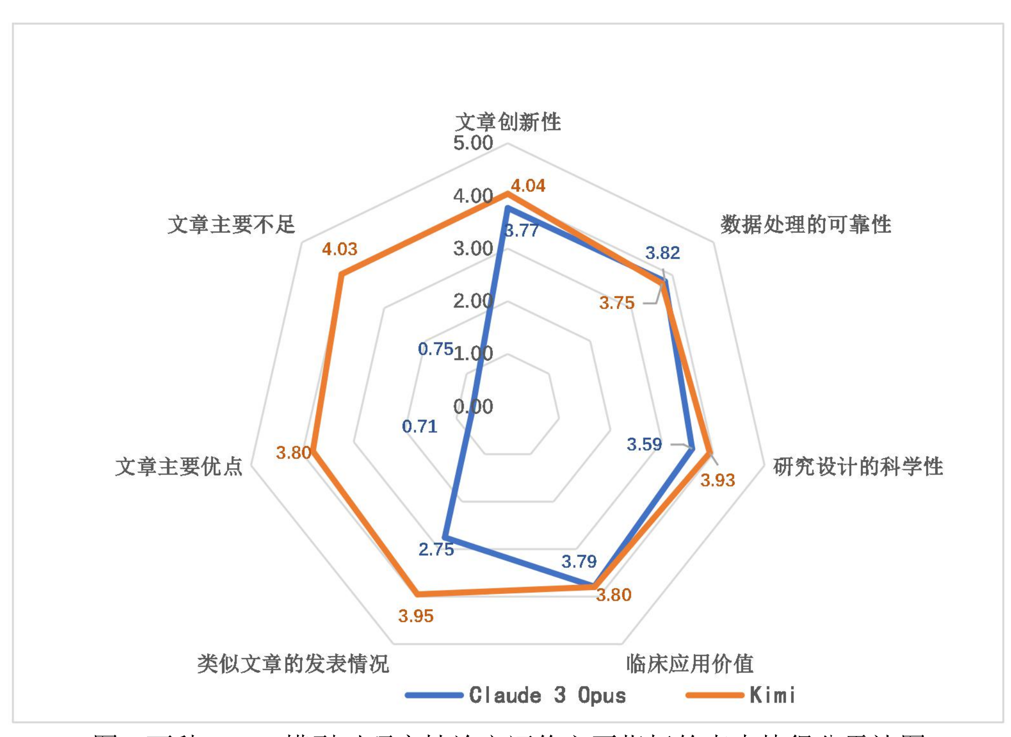


图 2 两种 AIGC 模型对研究性论文评价主要指标的李克特得分雷达图

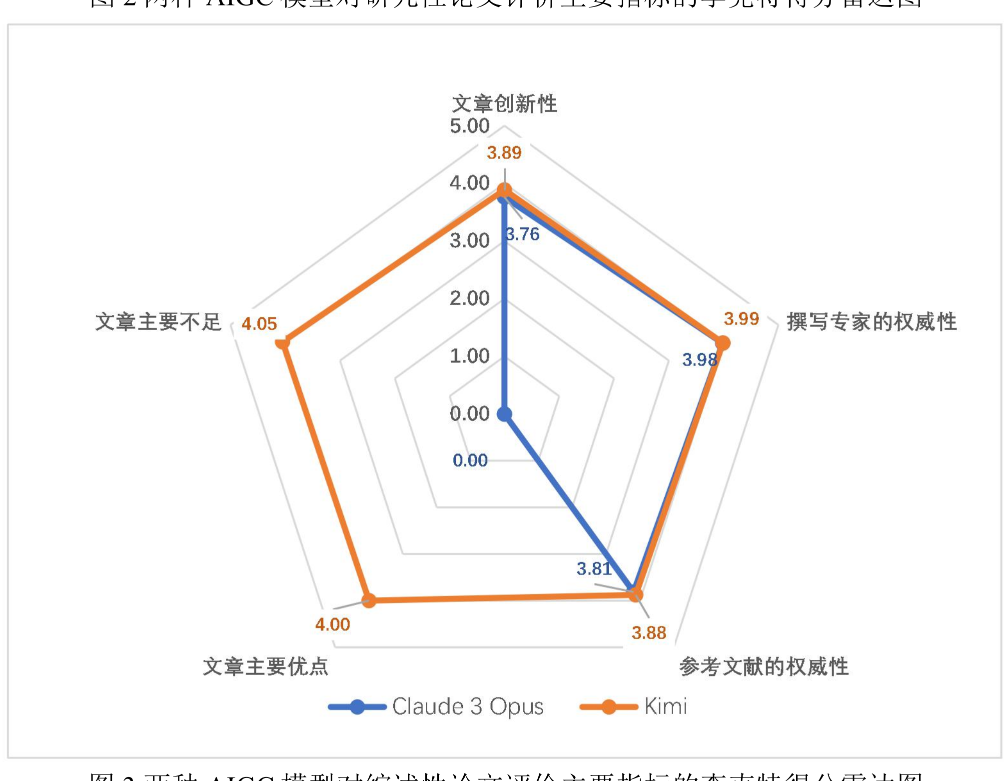


图 3 两种 AIGC 模型对综述性论文评价主要指标的李克特得分雷达图

2.4 不同文章类型披露情况的比较

本研究 29 篇文章按照对应《报告指南》，应披露清单共 710 条，除综述类文章外，责编、Claude 3 Opus 和 Kimi 对文章审查后的披露率分别为 53.7%、38.1%和 43.3%，经过 Person 卡方检验，差异有统计学意义（χ^2=99.532，P<0.001）。根据文章类型分层分析发现，动物实验类文章三者评审结果差异无统计学意义（见表 7）。进一步进行评审者两两卡方检验发现：与责编评价结果相比，Claude 3 Opus 评价结果在临床研究类文章（χ^2=4.601，P=0.331）、诊断实验类文章（χ^2=6.551，P=0.088）、病例报道类文章（χ^2=4.890，P=0.087）、

动物实验类文章（χ^2=5.942，P=0.051）方面，差异无统计学意义；而 Kimi 评价结果除了在观察性研究方面，差异有统计学意义外，在临床研究类文章（χ^2=5.855，P=0.119）、指南共识类文章（χ^2=2.357，P=0.308），诊断实验类文章（χ^2=6.153，P=0.104）、预测模型类文章（χ^2=5.489，P=0.064），病例报道类文章（χ^2=1.467，P=0.480）、动物实验类文章（χ^2=0.130，P=0.937）方面，差异无统计学意义。9 篇综述类文章，Claude 3 Opus 和 Kimi 按照 Prisma 披露，分别给出了 115 条和 198 条修改意见，占应披露率的 54.0%和 83.5%。

表 7 三者评价披露情况的皮尔森卡方检验结果

文章类型	比对结果	评审者			x^2	P值
		责编	Claude3	Kimi		
临床研究	披露	11	8	7	17.659	0.024
	部分披露	6	4	3		
	未披露	9	13	8		
	不适用	1	0	0		
	未评价	0	2	9		
指南共识	披露	59	39	50	54.405	<0.001
	部分披露	4	10	4		
	未披露	29	22	38		
	不适用	0	0	0		
	未评价	0	21	0		
观察性研究	披露	109	81	72	97.576	<0.001
	部分披露	12	35	34		
	未披露	36	52	41		
	不适用	19	6	0		
	未评价	0	2	29		
诊断试验	披露	19	14	20	19.572	0.003
	部分披露	1	6	0		
	未披露	10	8	4		
	不适用	0	0	0		
	未评价	0	2	6		
预测模型	披露	42	35	35	76.714	<0.001
	部分披露	20	20	34		
	未披露	46	22	39		
	不适用	0	1	0		
	未评价	0	30	0		
病例报道	披露	6	1	9	10.448	0.034
	部分披露	4	7	2		
	未披露	3	5	2		
	不适用	0	0	0		
	未评价	0	0	0		
动物实验	披露	5	0	5	6.819	0.146
	部分披露	9	10	10		
	未披露	7	11	6		
	不适用	0	0	0		
	未评价	0	0	0		
综述类文章	披露	24	20	6	703.081	<0.001
	部分披露	5	0	0		
	未披露/给意见	214	115	198		
	不适用	0	10	0		

未评价	0	98	39

2.5 2 种 AIGC 模型和责编评价对文章披露情况审查结果之间的比较

采用 Friedman *M* 检验对三者披露情况审查结果进行统计检验发现，Claude 3 Opus 的审查结果与责编（*M*=-0.314，调整后 *P*<0.001）和 Kimi（*M*=0.162，调整后 *P*=0.040）审查结果相比，差异有统计学意义，说明两组间审查结果见有差异；Kimi 的审查结果与责编审查结果之间，差异无统计学意义（*M*=-0.152，调整后 *P*=0.061），说明 2 者对披露清单的审查结果无差异，具有一致性。

2.6 2 种 AIGC 模型披露清单情况评价准确性的比较结果

以责编对文章披露情况作为金标准，将“部分披露”和“披露”定义为“披露”，将“未披露”、“不适用”定义为“未披露”。将 2 种 AIGC 评价结论的“披露”和“部分披露”定义为“阳性”结果，将“未披露”、“不适用”、“未评价”定义为“阴性”结果。采用四格表法对 Claude 3 Opus 和 Kimi 对披露清单评价与责编评价结果进行比较，结果显示 Claude 3 Opus 评价的准确性为 77.51%，灵敏度为 76.9%，特异度为 64.0%，阳性预测值为 87.4%，阴性预测值为 78.8%。Kimi 评价的准确性为 75.16%，灵敏度为 77.5%，特异度为 70.1%，阳性预测值为 83.5%，阴性预测值为 62.1%（表 8）。ROC 曲线分析结果显示，Claude 3 Opus 和 Kimi 曲线下面积分别为 0.818 和 0.841，有较好的预测能力（图 4）。

表 8 四格表法分析结果

AIGC 模型		责编		合计
		披露	未披露	
Claude 3 Opus	阳性	236	34	270
	阴性	71	126	197
Kimi	阳性	238	47	285
	阴性	69	113	182
合计		307	160	467

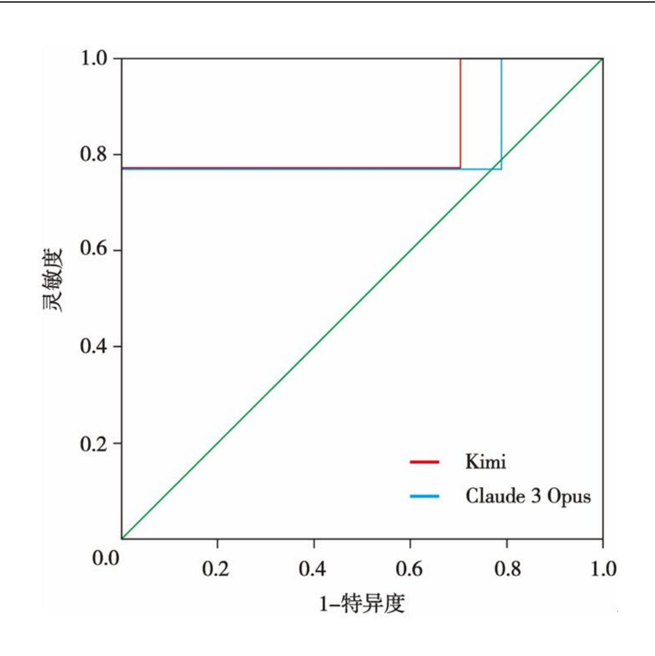


图 4 2 种 AIGC 审核结果的 ROC 曲线

3 讨论与分析

3.1 AIGC 在生物医学期刊同行评议中有一定应用前景，但评价能力尚需提升

随着 AI 技术的飞速发展和更新迭代，理解和对话能力不断提升，AI 的应用场景也不断扩大，也吸引了更多的关注和投入。笔者在日常整理专家同行评议意见中，也发现并证

实有专家采用了 AIGC 来进行同行评议，而且专家调研结果一致表示 AIGC 在同行评议中有应用前景。有研究结果显示，AIGC 应用于同行评议，有以下优势：①大幅提高效率[21,24-25]。AIGC 能快速处理大量数据，并准确地归纳论文核心内容呈现给同行评议专家，辅助其对论文进行评审；②降低同行评议偏见[25-26]。AI 可以避免因同学科专家之间的竞争或者个人偏见导致的不公正评议，能提供相对一致、公平、无偏见的评审意见；③文章质量把关[25-26]。AI 能帮助识别研究质量，发现错误，能按照期刊的标准来进行评价并反馈，同时也能提供语言清晰度、逻辑关联性等评价，协助提供改进文章质量的意见。

本研究结果发现，同行评议专家对 AIGC 同行评议的意见处于一般和同意之间，且检验发现，Kimi 与定稿会专家同行评议结论差异无统计学意义，但 Glaude 3 Opus 与定稿会专家同行评议结论差异存在显著性，说明 AIGC 在期刊同行评议方面的能力尚需要进一步的提升，在 2 者回答我们的问题时，还存在以下问题。①生成内容的不准确性。本研究在多次连续向 Claude 3 Opus 提问后，有 3 篇论文的同行评议结果出现了“答非所问”，应用于同行评议时，需要重点注意。同时，也有研究表示 AI 生成的同行评议内容的不准确[15,25]。可能的原因是提示词 1 中“创新性”“新颖性”“启发性”等提示词过于主观，导致，今后的研究中，需要进一步优化提示词，使 AIGC 的回答更为精准。②缺乏深度和个性化。主要表现在同行评议结论主要以归纳论文内容为主，同行评议结论表述趋于一致性，缺乏深入的文章内容的理解和推演，可能的原因是目前 AIGC 无法深入理解文章内容，缺乏人类审稿人的深度分析和个性化反馈[22,25]。③外部内容的推演和结合能力不足，尤其是针对医学领域的最前沿内容，因为没有足够的时间和最新数据的训练，往往无法给出准确的评价意见。④评审结论的模糊性。主要表现在 AIGC 给出的审稿结论较为模糊，通常是“前肯定后建议”的模式，给出退稿建议的底气不足，鲜有给出能击中文章致命弱点的退稿意见。这样的同行评议结论，最大的问题是导致编辑无法给出合适的返修或者退稿意见。⑤评审结论过于粗况，AIGC 有数次同行评议意见为：“目前尚未见有类似研究在国内外权威杂志上发表，该文章在研究内容和方法上具有一定的创新性。”“文章设计合理，实验方法恰当，对照组设置合理，研究思路清晰，层次分明，具有较强的科学性。”类似的同行评议内容，属于无效同行评议意见，需要重新评审。出现这种情况可能的原因是提示词不当，AIGC 算法不佳等。⑥审稿结论偏于正向评价。我们的研究中，AIGC 同行评议结论退稿少数，尤其是 Claude 3 Opus 评审结果，无一退稿，可能的原因是 AIGC 对论文进行同行评议仅仅是提取文章内的信息，如果文章内的内容以正向为主，则很可能得出的结论偏正向。

3.2 AIGC 在辅助生物医学期刊审查披露清单方面有一定优势

生物医学学术论文相较于其他学术文章，涉及到人体、动物、细胞等等，有其特殊性和复杂性。鉴于此，EQUATOR 网络于 2006 成立，制定了一系列的《报告指南》，并倡议通过这些规范，来提高生物医学研究的透明度，减少发表偏倚，提高研究的可重复性和可靠性，从而提高生物医学研究报告的质量。目前国际上一些知名的期刊，如《柳叶刀》《新英格兰医学杂志》《自然》《科学》等也鼓励或者要求作者在提交研究报告时遵循相应的 EQUATOR 规范。通常完整地按照《报告指南》中的披露清单进行撰写的文章被认为拥有较好的科学性、较高的可重复性和可信度。本研究结果显示，AIGC 模型对论文披露清单的审查，与责编的审查结果相比，有较高的准确性和一致性，且 AUC 均超过 0.8，且 AIGC 的审查迅速基本都在 2 分钟内完成并反馈结果，能高效地辅助责编完成对论文披露清单的审查。本研究采用现有的 AIGC 模型对论文披露清单进行审查，也未对 AIGC 模型进行特殊的训练和学习，也未对提示词进行非常系统化的斟酌，相信经过系统性训练的生物医学发表指南 AIGC 模型，有望能进一步提高审查的准确性，并成为生物医学期刊编辑身边提能增效的重要辅助工具。

4 AIGC 应用于生物医学期刊同行评议的建议

目前大部分期刊不接受 AIGC 用于同行评议，小部分期刊要求专家使用 AIGC 进行同行评议时，需要明确注明[27-28]。AIGC 技术正以迅猛的速度发展，尽管目前它还不能完全独立地承担生物医学期刊的同行评议任务，但在辅助编辑和专家评审论文的特定场景中，已经展现出其独特的优势和巨大潜力。虽然还存在诸多的缺陷和不足，但是 AIGC 的应用于同行评议的前景，为了进一步拓展 AIGC 在生物医学期刊同行评议领域的应用，以下是一些需要优化的关键方面：

4.1 明确责任和制定规范

在传统的同行评议中，评审者对其评审结果负有直接责任。但当 AI 介入后，它作为一个工具，其决策过程可能缺乏足够的透明度，如果出现错误或争议，缺乏明确的问责机制来追究责任，使得责任归属变得模糊[25]。如果评审结果出现问题，难以确定是 AI 算法的问题，还是使用 AI 的评审者的问题。笔者认为使用 AIGC 辅助审稿的专家，有义务对 AICG 生成的同行评议结果进行把关，理应为 AIGC 的同行评议结果负责。同时，随着 AIGC 技术的蓬勃发展，确立相应的规则和标准对于其健康有序的发展至关重要。国家网信办、发展改革委等[29]7 部委于 2023 年 8 月发布了《生成式人工智能服务管理暂行办法》，鼓励 AIGC 技术在各行业、各领域的创新应用，生成积极健康、向上向善的优质内容，探索优化应用场景，构建应用生态体系。同时也对 AIGC 服务提供者、数据质量、知识产权、信息安全、内容质量提出了要求，坚持发展和安全并重，进一步为 AIGC 服务在各领域中的应用提供了明确的规范和指导。中国音像与数字出版协会[30]于 2024 年 1 月发布实施《出版业生成式人工智能技术应用指南》团体标准，对 AIGC 应用的基本原则、主要应用场景、管理机制、知识产权、安全保障等方面进行了规定。中国科学技术信息研究所联合爱思唯尔、斯普林格·自然、约翰威立国际出版社集团在征求各方意见的基础上，制定了《学术出版中 AIGC 使用边界指南》，明晰了相关利益主体在学术期刊论文准备、写作、投稿、评审、出版、传播各环节应该履行的最佳行为实践，提供规范的 AIGC 使用指导，以期达到防范学术不端、加强诚信治理和引导相关利益主体就使用 AIGC 达成共识的目的[31]。目前，尚未有专门针对 AIGC 应用于生物医学期刊同行评议的使用指南。为此，需要生物医学专家、医学编辑、AI 技术专家和伦理学家等多方共同参与，制定明确的规则和标准。这包括明确 AIGC 在同行评议中的作用和限制，规定使用范围、操作流程，以及在评议过程中的决策权重等，以确保 AIGC 技术的合理、有效和安全应用。

4.2 关注评审透明度

AI 的算法通常被视为"黑箱"，其使用的训练数据、内部工作机制和决策逻辑对外部人员来说不可见[19,26,32]。这种不透明性可能存在潜在偏倚，导致对 AI 同行评议结果的信任度下降。评审透明度是确保 AIGC 技术在同行评议中公正性和有效性的关键，主要有以下几个方面：①增加数据的透明度：公开 AIGC 模型所依赖的数据来源，详细说明数据集的类型、规模、收集方法和时间范围。数据透明度的提高有助于确保数据的可靠性和代表性，为评审者和作者提供清晰的数据背景；①开放算法透明度，使评审者和作者能够理解 AIGC 的决策逻辑，不仅增强了模型的可靠性，也提高了其可信度，使得评审过程更加透明和可解释。③提高 AIGC 输出结果的透明度：输出结果不应仅限于单一结论，而应包含详细的分析和依据，以便编辑和出版人员能够全面评价和审核 AIGC 的输出结果的正确性，让编辑和评审者能够对 AIGC 的评审结果进行深入的理解和恰当的运用。

4.3 保证数据的权威性和时效性

生物医学研究日新月异，可能会存在 AIGC 的学习和训练不及时问题，可能缺乏对文章核心问题的深入分析[3,22]。本研究中针对提示词中"是否有类似文章的发表情况"，KIMI 的答复仅仅针对被评审文章的参考文献进行了检索，并未能检索常用的外部数据库；Claude 的答复多数以"目前尚未见类似研究报道"为主。如何对 AIGC 模型进行及时的训

练，让其获取科学的前沿生物医学内容是保证 AIGC 同行评议的准确性的根本。AIGC 在生物医学期刊同行评议中的准确性，根本上取决于其学习数据的权威性和时效性。可以采取以下措施：①建立权威数据集：开发和构建专门针对生物医学期刊同行评议的高质量数据集，这些数据集应来源于公认的权威机构或经过同行评议的学术出版物。②保持数据更新：定期更新数据集，纳入最新的研究成果和行业动态，以确保 AIGC 能够反映最新的知识和学术前沿。③数据预处理与清洗：对数据进行严格的标注、预处理和清洗，以消除错误、偏见和冗余信息，确保数据集的准确性、多样性和权威性。④实施数据质量监控：建立持续的数据质量监控机制，对数据集进行定期评估和维护，以适应不断变化的学术研究需求。⑤促进数据共享与合作：鼓励有关科研机构、权威生物医学数据库和 AIGC 开发者之间的数据共享与合作，以丰富数据集内容，提高数据的覆盖面和深度。

4.4 注重相关医学伦理和数据安全

用 AIGC 进行同行评议的学习、训练和验证中，可能会涉及患者隐私、人群偏见、临临床试验伦理、医学伦理等内容，可能存在潜在的数据泄露和出版伦理问题，主要表现在①未发表数据泄露问题；②患者隐私泄露问题；③在特定场景和蓄意诱导下，可能出现虚构结论问题[15,22,33]。需要强化规则，设计更有效的监管策略来预防和消除这些问题的发生。笔者认为，一旦论文录用后，可以让作者先上传至预出版平台，再用 AIGC 工具进行同行评议，这样可以规避数据泄露和出版伦理问题。另外，除了要增加评审透明度以外，还需要在 AI 设计和研发时嵌入相关法律法规和医学伦理规范，并对 AI 系统进行伦理审查，确保其设计和应用符合医学相关伦理标准。同时，期刊也应明确声明 AIGC 用于同行评议相关场景、规范和原则等，并加强对 AIGC 同行评议结论的监督审查。

4.5 建立监督反馈机制

目前 AIGC 的同行评议结果在准确性尚存在一定偏倚，也无法取代人工同行评议结果。故采用 AIGC 进行同行评议后，应由科学编辑或生物医学专家进行复核，确保评议的准确性和公正性。同时，需要建立对 AIGC 评议结果的反馈机制，以便不断优化和改进 AIGC 的同行评议模型。

4.6 加强培训教育

加强对 AIGC 使用者的教育培训，使他们了解 AIGC 的优势和局限性，提高其对 AIGC 应用能力和驾驭能力，提升对 AIGC 的认知以及如何在评议中合理利用 AIGC 技术。

5 结语

AI 技术的迅猛发展带来了日新月异的变化，AIGC 模型的应用领域也在不断拓宽，其更新迭代周期显著缩短。本研究揭示了 Claude 3 Opus 和 Kimi 应用于生物医学期刊同行评议中虽尚存在一些不足之处，但已展现出一定的潜力，尤其是 Kimi 的评议结果与专家定稿的评议结果在统计学上并无显著差异。李克特评分结果也显示专家们对这两种 AIGC 模型的评议结论的认可度处于一般到同意之间，但他们均表示未来 AIGC 技术有望在同行评议中发挥重要作用。两种 AIGC 模型在审查《报告指南》的规范披露清单目时，与责任编辑的审查结果在统计学上没有显著差异，AUC 曲线下面积均超过 0.8，审查的准确的都超过了 75%，显示出较高的准确性，同时也提高的工作效率，展现出了巨大的优势。尽管如此，AIGC 技术目前还无法完全取代人工同行评议。本研究还存在提示词过于主观性，前期训练不足、提问重复次数不多、不同的研究类型纳入文献样本不多等问题，需要更进一步的扩大样本，优化提示词来测试 AIGC 在生物医学期刊同行评议中的能力。当然，我们仍需关注 AIGC 应用于同行评议中的使用边界、评审偏倚、数据安全、出版伦理以及训练的时效性等问题，也相信这些问题在开放科学的发展、人工智能技术的迭代下，会逐一进行优化和解决。我们也要呼吁管理者、审稿专家、医学编辑和 AIGC 开发者等相关人员以更加积极的态度来迎接人工智能时代的到来，进一步加强合作，共同推动顶层制度的建设，提高数据和算法的

透明度，提升数据的时效性和权威性，注重出版伦理，建立有效的监督反馈机制，以确保AIGC 产品的性能，并提升其在同行评议中的能力。随着技术的不断进步，我们有理由相信，AIGC 技术必将成为编辑和同行评议专家身边的重要工具。

5 致谢

感谢复旦大学附属肿瘤医院杂志社办公室王琳辉副编审在本研究中审核论文披露清单中所做的贡献。

Evaluation of Claude 3 Opus and Kimi in peer review for biomedical journals

NI Ming

Department of Editorial Office, Fudan University Shanghai Cancer Center, Department of Oncology, Shanghai Medical College, Fudan University, 270 Dong'an Road, Shanghai 200032, China

Abstract [Purposes] This study aimed to evaluate the capabilities of Claude 3 Opus and Kimi in peer review for biomedical journals. **[Methods]** Peer reviews were conducted using Claude 3 Opus and Kimi on 29 papers from *China Oncology* that were at the final review stage before publication, and were examined according to the disclosure checklist in the Biomedical Research Reporting Guidelines. All authors declared their consent to use AIGC for peer review. The results of the AIGC peer reviews were scored using a 5-point Likert scale. Count data were analyzed using one-way ANOVA or *Fisher*'s exact probability test, and multiple paired measurement data were analyzed using the Friedman *M* test. Sensitivity, specificity, positive predictive value, negative predictive value, and accuracy of the two types of AIGC assessments were calculated using a four-fold table method, and ROC curves were plotted and assess their predictive capability. **[Findings]** Of the 29 articles, after the final review meeting, 6 were published, 15 were published after revisions, and 8 were rejected. Claude 3 Opus's peer review conclusions led to the publication of 19 articles and the publication of 10 after revisions. Kimi's peer review conclusions resulted in the publication of 9 articles, 16 after revisions, and 4 rejections. The Friedman *M* test showed no statistically significant difference between Kimi's peer review conclusions and the experts' conclusions at the final review meeting (M=0.241, adjusted P=1.000). The Likert scale results indicated that the experts at the final review meeting had a higher level of agreement with Kimi's peer review results than with those of Claude 3 Opus (3.85±0.47 *vs* 3.48±0.73, F=10.017, P=0.002). In terms of reviewing according to the Biomedical Research Reporting Guidelines disclosure checklist, Claude 3 Opus's evaluation accuracy was 77.5%, sensitivity was 76.9%, and specificity was 64.0%; Kimi's evaluation accuracy was 75.2%, sensitivity was 77.5%, specificity was 70.1%. ROC curve analysis showed that the areas under the curve for Claude 3 Opus and Kimi were 0.818 and 0.841, respectively, indicating good predictive capability. Upon testing, there was no statistically significant difference between Kimi's review results and the editor's review results (M=-0.152, adjusted P=0.061). **[Conclusions]** Claude 3 Opus and Kimi have shown good capabilities in reviewing the disclosure checklist of biomedical research reports, with high consistency compared to editorial reviews. However, these two AIGC models have not yet reached the level of expert peer review, exhibiting issues such as: ① Inaccuracy in generated content; ② Lack of personalization in generated content; ③ Insufficient extrapolation of external

content; ④ Ambiguity in review content; ⑤ Overly coarse-grained generated content; ⑥ Bias towards positive evaluations in review conclusions. Despite these issues, they still demonstrate certain potential for application. To further enhance their effectiveness, it is recommended that medical experts, medical editors, and AIGC developers work together to establish relevant standards, ensure data security and quality, increase transparency, reduce review bias, adhere to publishing ethics, and establish effective supervision and feedback mechanisms to ensure the accuracy of AIGC in peer review for biomedical journals.
Key words: Biomedical journals; Generative artificial intelligence; Claude 3 Opus; Kimi; Peer review

参考文献

[1] EYSENBACH G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers［J］. *JMIR Medical Education*, 2023, 9: e46885.
[2] ELSE H. Abstracts written by ChatGPT fool scientists［J］. *Nature*, 2023, 613(7944): 423.
[3] 刘芳. 生成式人工智能快速发展时代科技期刊编辑价值升维策略[J]. 编辑学报, 2023, 35(S1): 111-116.
[4] 王萌, 李瑜, 陈昊旻, 等. ChatGPT 在辅助写作过程中的版权归属问题探讨[J]. 编辑学报, 2023, 35(S2): 133-134.
[5] 崔玉洁. ChatGPT 与人工编校相结合: 提高期刊编校效率和文章质量[J]. 编辑学报, 2023, 35(4): 429-433.
[6] 陈小明. 生成式人工智能技术对医学科技期刊学术生产与出版的影响[J]. 编辑学报, 2023, 35(S1): 132-136.
[7] 郑莉, 刘惠琴. 人工智能浪潮下科技期刊编辑的必备技能及能力提升策略［J］. 天津科技, 2024, 51(1): 85-89.
[8] 沈锡宾, 王立磊. 人工智能生成学术期刊文本的检测研究[J]. 科技与出版, 2023(8): 56-62.
[9] 张璐, 郭晓亮, 景勇, 等. ChatGPT 对学术期刊的影响及应对策略研究［J］. 出版与印刷, 2023(4): 91-96.
[10] 张重毅, 牛欣悦, 孙君艳, 等. ChatGPT 探析: AI 大型语言模型下学术出版的机遇与挑战[J]. 中国科技期刊研究, 2023, 34(4): 446-453.
[11] 黄伟, 孙伟, 蒋霞. “互联网+” 时代下同行评议期刊精准寻找审稿专家的途径与实践［J］. 编辑学报, 2023, 35(1): 60-65.
[12] 杨雅. 生成式人工智能在科技期刊出版中的应用场景探讨［J］. 新闻研究导刊, 2024, 15(2): 242-245.
[13] LIANG W X, IZZO Z, ZHANG Y H, et al. Monitoring AI-Modified Content at Scale: A Case Study on the Impact of ChatGPT on AI Conference Peer Reviews[EB/OL]. (2024-06-15)[2024-08-05]. https://arxiv.org/abs/2403.07183.
[14] SINGH CHAWLA D. Is ChatGPT corrupting peer review? Telltale words hint at AI use［J］. *Nature*, 2024, 628(8008): 483-484.
[15] 袁庆, 沈锡宾, 刘红霞, 等. 中国科技期刊编辑大模型技术认知及其影响的调研研究［J］. 编辑学报, 2024, 36(2): 183-188.
[16] 董文杰, 李苑. 人工智能在科技期刊中的应用及启示［J］. 中国科技期刊研究, 2023, 34(11): 1399-1408.
[17] 张彤, 唐慧, 胡小洋, 等. 人工智能辅助学术期刊同行评议的功能需求分析［J］. 编辑学报, 2021, 33(5): 523-528.
[18] 江雨莲, 孙激. 人工智能在医学期刊编辑出版中的应用［J］. 科技与出版, 2020(2): 66-71.
[19] PERCHIK J D, SMITH A D, ELKASSEM A A, et al. Artificial intelligence literacy: Developing a multi-institutional infrastructure for AI education［J］. *Academic Radiology*, 2023, 30(7): 1472-1480.
[20] KADI G, ALI ASLANER M. Exploring ChatGPT’s abilities in medical article writing and peer review［J］. *Croatian Medical Journal*, 2024, 65(2): 93-100.

[21] YORUKOGLU K. Can artificial intelligence (AI) be a reviewer of a medical article? ［J］. *Türk Patoloji Dergisi*, 2024, 40(2): 75-77.
[22] Ertugrul P. Advantages of Claude 3: Is claude better than GPT-4?[EB/OL]. (2024-03-19)[2024-06-05]. https://textcortex.com/post/advantages-of-claude-3.
[23] 晒科网. 【大模型最新排行】Claude3 称王，Elo 超越 GPT4 最新版，通义千问进前 10[OL]. (2024-03-28)[2024-12-22].https://zhuanlan.zhihu.com/p/689562075.
[24] ZHANG J. Kimi 大模型激起“鲇鱼效应”，长文本卷起新高度[OL]. (2024-03-26)[2024-12-22]. https://www.eet-china.com/news/202403263912.html.
[23] EQUATOR. Reporting guidelines for main study types[EB/OL]. [2024-07-05].https://www.equator-network.org/.
[24] MEHTA V, MATHUR A, ANJALI A K, et al. The application of ChatGPT in the peer-reviewing process［J］. *Oral Oncology Reports*, 2024, 9: 100227.
[25]胡靖宇, 杨博, 高海军. 科技期刊对 ChatGPT 冲击的应对策略[J]. 编辑学刊, 2024, (1):80-85.
[26] BAUCHNER H, RIVARA F P. Use of artificial intelligence and the future of peer review［J］. *Health Affairs Scholar*, 2024, 2(5): qxae058.
[27] CHENG K M, SUN Z J, LIU X J, et al. Generative artificial intelligence is infiltrating peer review process［J］. *Critical Care*, 2024, 28(1): 149.
[28]洪悦民,王景周.学术期刊出版中人工智能生成内容的使用规范及著录建议[J].编辑学报,2024,36(2):149-153.
[29] 国家互联网信息办公室, 中华人民共和国国家发展和改革委员会, 中华人民共和国教育部, 等. 生成式人工智能服务管理暂行办法[S]. 国务院公报, 2023(24): 39-42.
[30] 中国音像与数字出版协会. 出版业生成式人工智能技术应用指南[S/OL]. (2023-12-20)[2024-07-05]。http://www.cadpa.org.cn/3281/202312/41644.html.
[31] 中国科学技术信息研究所. 学术出版中 AIGC 使用边界指南[S/OL]. (2023-09-22)[2024-07-05]. https://www.istic.ac.cn/html/1/245/1701698014446298352.html.
[32] GOMES W J, EVORA P R B, GUIZILINI S. Artificial intelligence is irreversibly bound to academic publishing - ChatGPT is cleared for scientific writing and peer review［J］. *Brazilian Journal of Cardiovascular Surgery*, 2023, 38(4): e20230963.
[33] 雷芳, 杜亮, 董敏, 等. 关于人工智能背景下医学期刊应对医学伦理问题的思考［J］. 编辑学报, 2023, 35(3): 263-267.

作者贡献声明：
倪明：研究设计、问卷发放、数据整理分析、论文撰写